Gerald Amiel Ballena

gmballena@up.edu.ph github.com/GABallena linkedin.com/in/gerald-amiel-ballena
Metro Manila, Philippines orcid.org/0009-0000-8857-9755

Professional Summary

Versatile bioinformatics specialist with extensive experience in developing scalable workflows, computational biology, and machine learning. Successfully processed multi-terabyte datasets, streamlined bioinformatics pipelines to enhance reproducibility, and enabled actionable insights in environmental and public health projects.

Experience

2024 – Present **Project Technical Specialist**
University of the Philippines, College of Public Health

- Developed scalable bioinformatics workflows for high-throughput sequence analysis (2.5 Terabytes each).
- Automated data preprocessing workflows using Snakemake, reducing manual load by 90% when data is available while also enhancing reproducibility and removing human error.
- Collaborated with cross-disciplinary teams on projects involving public health, microbiology, and science of environmental engineering.
- Enhanced data analysis pipelines, leading to actionable insights for surveillance and public health research.

Education

University of the Philippines Diliman

Graduated: July 2022

Thesis: In silico assessment of the association of pathogenicity and metal-resistance potential of *Fusarium* spp. Pre-print Link

Accomplishments: DOST ASTHRDP-Scholarship
Calculated Effective GPA (MS): 1.72

Relavant Courses

- Industrial microbiology
- Molecular Phylogenetics
- Epigenetics
- Advanced Cell and Molecular Biology

University of the Philippines Baguio

Graduated: June 2018

Thesis: **Bioelectrocatalysis by Novel Electrogenic Alkaliphilic Bacteria *Bacillus* sp. BAB-3442 Using Dual-Chambered Microbial Fuel Cell** ..... Poster Presentation (PSM 47)

Accomplishments: Advanced Placement Exam: Advanced Algebra

Philippine Science High School CAR Campus

Graduated: March 2014

Accomplishments: Focused on STEM curriculum with a strong emphasis on research and scientific inquiry.

References available only upon request and if shortlisted

Skills

Technical Skills

Programming Languages: Python (Fluent), R (Advanced), Bash (Intermediate), Perl & BioPerl (Intermediate), C++ (Fair).

Bioinformatics Tools: Kraken2, MetaPhlAn, MEGAHIT, Snakemake, Prokka, Jellyfish, BUSCO, and others.

Workflow Automation: Snakemake, Conda, YAML (Configuration Files).

Data Visualization: ggplot2, Plotly, matplotlib, QGIS.

High-Performance Computing: Slurm, Docker.

Soft Skills

Collaboration: Proficient in leading cross-disciplinary projects and fostering effective team collaboration.

Technical Writing: Skilled in preparing technical reports and documentation using TeX tools such as TeXStudio and Overleaf.

Problem-Solving: Adept at diagnosing and optimizing bioinformatics pipelines to enhance efficiency, reproducibility, and both statistical and scientific robustness.

Certifications and Relevant Coursework

Certifications

Data Analyst Associate: DataCamp 2024

AI Fundamentals: DataCamp 2024

Data Literacy: DataCamp 2024

Principles, Statistical and Computational Tools for Reproducible Data Science: Harvard EdX . . 2024

Relevant MOOCs

Introductory: Introduction to Data Engineering 2020

Introductory: COVID-19 Contact Tracing 2020

Introductory: Mind Control: Managing Your Mental Health During COVID-19 2020

Introductory: COVID-19: What You Need to Know 2020

Introductory: Essential Epidemiologic Tools for Public Health Practice 2020

Intermediate: Supervised Learning with sci-kit learn 2024

Intermediate: Visualizing GeoSpatial Data in `Python` 2024

Intermediate: Working with Geospatial Data in `Python` 2024

Intermediate: Unsupervised learning in `Python` 2024

Advanced: Ensemble Methods in `Python` (Bagging, Boosting, Stacking) 2024

Intermediate: Biostatistics in Public Health 2020

- Summary Statistics in Public Health
- Hypothesis Testing In Public Health
- Simple Regression Analysis in Public Health
- Multiple Regression Analysis in Public Health

Intermediate: Genomic Data Science 2024

- Introduction to Genomic Technologies
- Python for Genomic Data
- Tools for Genomic Data Science
- Command line tools for Genomic Data Science

Intermediate: Genomic Analysis track 2024

References available only upon request and if shortlisted

- Bioconductor in R
- RNA-Seq with Bioconductor
- Differential Expression Analysis with limma
- CHIP-Seq with Bioconductor in R

Intermediate: Visualizing Geospatial Data in R 2024
Advanced: Hyperparameter Tuning in R, Designing Machine Learning Workflows in Python 2024
Advanced: Bioinformatics via Coursera 2024
(*University of California, San Diego*)
- Finding Hidden Messages (with Honors)

Professional Development

- Introduction to AWS

Scripts and Workflows

Key Pipelines and Workflows: Developed and implemented scalable workflows and pipelines using Snakemake, Python, and Bash for metagenomic analysis, diversity profiling, and bioinformatics tool management. Highlights include:

- **Metagenomic Analysis Pipeline:** Automated workflows for trimming, taxonomic profiling (**Kraken2**, **Bracken**), and diversity calculations (**scikit-bio**), ensuring high reproducibility and scalability.
- **Comprehensive Binning Workflow:** Designed workflows for assembly, binning (**MetaWRAP**, **MEGAHIT**), and MAG validation (**CheckM2**), streamlining large-scale metagenomic projects.
- **Plasmid and ARG Analysis:** Built pipelines for plasmid detection and antimicrobial resistance profiling using **metaSPAdes**, **PlasmidFinder**, and **RGI**.
- **Diversity Analysis:** Developed Python-based scripts and R workflows for alpha/beta diversity metrics (Shannon, Chao1, Bray-Curtis) and visualizations using **ggplot2**.

Automation and Custom Tools: Created tools and workflows for parameter optimization, repository mining, and bioinformatics tool management:

- **Randomized Parameter Testing:** Automated preprocessing parameter exploration for tools like **Trimmomatic**, **Cutadapt**, and **fastp**, enabling systematic optimization.
- **Bioconda Repository Mining:** Developed a Python-based scraper to extract and filter bioinformatics tools for metagenomics and AMR research.
- **General Bootstrapping Workflow:** Automated sampling of paired-end reads for diversity and functional analyses using **seqtk**.
- **Tool Management:** Streamlined dependency discovery, YAML updates, and Conda-based environment management for reproducible pipelines.
- **Statistical Optimization for k-mer Profiles:** Designed a Snakemake workflow integrated with custom Python scripts to identify the best statistical fit (non-parametric, parametric, or transformed parametric) for k-mer profiles, facilitating the selection of appropriate statistical tests for raw data analysis.

Advanced Analysis and Statistical Workflows: Designed workflows for k-mer analysis, contaminant detection, and statistical evaluations:

- **K-mer Analysis:** Automated frequency distribution fitting, entropy calculations, and alignment validation for metagenomic datasets (**Jellyfish**, **MASH**).
- **Contaminant Filtering:** Built pipelines for k-mer mapping and statistical testing against known contaminant databases (**UniVec**, **PhiX**, **KMA**).
- **Visualization Pipeline:** Developed R-based workflows for ridgeline and violin plots, NMDS, and heatmaps to visualize diversity and taxonomic profiles.

References available only upon request and if shortlisted